MATH/STAT 3460, Intermediate Statistical Theory
Winter 2014
Toby Kenney
Sample Final Examination
Model Solutions

1. *Under a certain model, the number of is assumed to follow a censored Poisson distribution with parameter λ. That is the probabilites of 0, 1 and 2 are $e^{-\lambda}$, $\lambda e^{-\lambda}$ and $1-e^{-\lambda}(1+\lambda)$. The observed frequencies are :*

Number	*Frequency*
0	*102*
1	*131*
2	*84*

Use Newton's method to find the Maximum Likelihood estimate for λ. [Start with an initial estimate of 1, and perform 1 iteration.]

The log-likelihood is $l(\lambda) = -102\lambda - 131(\lambda - \log(\lambda)) + 84\log(1-e^{-\lambda}(1+\lambda))$. The score function is $S(\lambda) = -102 + \frac{131}{\lambda} - 131 + 84\frac{\lambda e^{-\lambda}}{1-e^{-\lambda}(1+\lambda)}$. The information function is $\mathcal{I}(\lambda) = -\frac{131}{\lambda^2} - 84\frac{(1-\lambda)e^{-\lambda}(1-e^{-\lambda}(1+\lambda)) - \lambda^2 e^{-2\lambda}}{(1-e^{-\lambda}(1+\lambda))^2} = -\frac{131}{\lambda^2} - 84\frac{(1-\lambda)e^{-\lambda} - (e^{-2\lambda}(1+\lambda)(1-\lambda)) - e^{-2\lambda}}{(1-e^{-\lambda}(1+\lambda))^2} = -\frac{131}{\lambda^2} + 84\frac{2\lambda e^{-\lambda} + e^{-2\lambda}}{(1-e^{-\lambda}(1+\lambda))^2}$.

We get the following

λ	$S(\lambda)$	$\mathcal{I}(\lambda)$
1	14.946	321.87
1.0464		

So after one step, we have an estimate of 1.0464.

2. *A scientist wants to determine the frequency of a particular version of a gene. The gene has two versions A (probability p) and B (probability $1-p$), and individuals have two copies of the gene, and so can be classified as AA (probability p^2), AB (probability $2p(1-p)$) or BB (probability $(1-p)^2$). The scientist has two tests — one that tests for the state AA and one that tests for the state BB. The scientist plans to test 100 patients with one of the two tests. If the true value of p is 0.3:*

(a) What is the expected information about p for each of the tests?

Let N be the number of patients identified by the test. The likelihood of the data is $L(p) = p^{2N}(1-p^2)^{100-N}$ for the first test, and $L(p) = (1-p)^{2N}(1-(1-p)^2)^{100-N}$ for the second test. The score functions are $\frac{2N}{p} - \frac{100-N}{1-p} + \frac{100-N}{1+p}$ and $-\frac{2N}{1-p} + \frac{100-N}{p} - \frac{100-N}{2-p}$ respectively. The information functions are $\frac{2N}{p^2} + \frac{100-N}{(1-p)^2} + \frac{100-N}{(1+p)^2}$ and $\frac{2N}{(1-p)^2} + \frac{100-N}{p^2} + \frac{100-N}{(2-p)^2}$.

For the first test, $\mathbb{E}(N) = 100p^2$, so the expected information is $200 + \frac{100(1+p)}{(1-p)} + \frac{100(1-p)}{(1+p)}$, which for $p = 0.3$ is 439.56. For the second test,

$\mathbb{E}(N) = 100(1-p)^2$, so the expected information is $200 + \frac{100(2-p)}{p} + \frac{100p}{2-p}$, which for $p = 0.3$ is 784.13.

(b) Which test should the scientist use?

The scientist should generally use the second test. [Though there might be other considerations which could change this.]

3. *Every year a certain city is flooded with probability p. Based on data from a record of whether or not the city flooded every year for the past 600 years. The maximum likelihood estimate for p is therefore $\hat{p} = \frac{F}{600}$, where F is the number of years that the city is flooded.*

   An insurance company is interested in the probability that the city floods at any time within the next 2 years. The probability of this is $1-(1-p)^2 = 2p-p^2$. The maximum likelihood estimate of this is therefore $\frac{F}{300} - \frac{F^2}{360000}$. What is the bias of this estimate?

   We have that $F \sim B(600, p)$, so $\mathbb{E}(F) = 600p$ and $\mathrm{Var}(F) = 600p(1-p)$, so $\mathbb{E}(F^2) = \mathbb{E}(F)^2 + \mathrm{Var}(F) = 360000p^2 + 600p(1-p)$. We therefore have $\mathcal{E}(\frac{F}{300} - \frac{F^2}{360000}) = \frac{600p}{300} - \frac{360000p^2+600p(1-p)}{360000} = 2p - p^2 - \frac{p(1-p)}{600}$. The bias of this estimate is therefore $-\frac{p(1-p)}{600}$.

4. *Let X_1, X_2, X_3 be samples from a Poisson distribution with parameter λ. Suppose $X_1 = 3$, $X_2 = 0$, and $X_3 = 4$. Find a 95% confidence interval for λ. [Hint: the endpoints of the interval are in the set $\{1.003, 2.001, 3.619, 4.513, 4.925\}$.]*

   The log-likelihood of the data is $l(\lambda) = -3\lambda + 7\log\lambda$. The score function is therefore $\frac{7}{\lambda} - 3$, and the maximum likelihood estimate for λ is $\frac{7}{3}$. Using a chi-squared approximation, a 95% confidence interval occurs when the relative log-likelihood is at least $-\frac{3.84145882}{2}$, so we need to solve $7\log\lambda - 3\lambda \geqslant -7 + 7\log\left(\frac{7}{3}\right) - \frac{3.84145882}{2} = -2.989644387$.

   We evaluate

λ	$7\log\lambda - 3 - \lambda$
1.003	−2.988
2.001	−1.147
3.619	−1.854
4.513	−2.990
4.925	−3.615

   We therefore find that the 95% confidence interval is $[1.003, 4.513]$.

5. *Let $X_1, \ldots, X_{20}$ be samples from a Poisson distribution with parameter λ. Suppose that $X_1 + \cdots + X_{20} = 123$. Using a likelihood ratio test, test the hypothesis that $\lambda = 7$ at the 5% significance level.*

   The log-likelihood is $123\log(\lambda) - 20\lambda$. The maximum likelihood estimate is therefore $\lambda = \frac{123}{20} = 6.15$. The relative log-likelihood of $\lambda = 7$ is therefore $123\log(7) - 140 - 123\log(6.15) + 123 = -1.07665773$. The chi-square statistic is -2 times this log-likelihood ratio, which is -2.1533. The p-value of this is 0.142, so the hypothesis is not rejected.

6. *Let $X_1, \ldots, X_{50}$ be samples from a Normal distribution with mean μ and variance σ^2. Suppose that the sample mean is 2.3 and the sample variance is 6.25. Using a likelihood ratio test, test the hypothesis that $\sigma = 3$ at the 5% significance level.*

   The log-likelihood is $-50\log(\sigma) - \sum \frac{(X_i-\mu)^2}{2\sigma^2} = -50\log(\sigma) - 50\frac{6.25+(2.3-\mu)^2}{2\sigma^2}$. For fixed *sigma*, this is maximised by $\mu = 2.3$. The maximum relative log-likelihood of $\sigma = 3$ is therefore $l(2.3, 3) - l(2.3, 2.5) = 50\log(2.5) - 50\log(3) + \frac{31.25}{6.25} - \frac{31.25}{9} = -7.588300062$. The chi-squared statistic is -2 times this, which is 15.177. There is one degree of freedom, so the p-value is 0.00009789, so the hypothesis is rejected.

7. *In a sample of 100 variables $X_1, \ldots, X_n$, it is believed that the X_i are independent samples from a binomial distribution with $n = 2$ and $p = 0.1$. The following results are obtained:*

Value	*0*	*1*	*2*
Frequency	*81*	*16*	*3*

   Test the hypothesis using a chi-squared test at the 10% significance level.

   The expected values are 81, 18 and 1. The chi-squared statistic is therefore $0 + \frac{2^2}{18} + \frac{2^2}{1} = 4.22$. This is compared to a chi-square distribution with 2 degrees of freedom, so it has a p-value of 0.1211, so it is not rejected.

   [The log-likelihood ratio is $18\log(0.18) - 18\log(0.16) + 3\log(0.01) - 3\log(0.03) = -1.176$, so a likelihood ratio test has a chi-square statistic of 2.351, which gives a p-value of 0.308, so the chi-squared approximation is not very accurate in this case.]

8. *A scientist is studying the effects of vitamin supplements on intelligence. She takes 1000 subjects, and gives vitamin supplements to 500 of them for a period of 3 months. Then she gives them all a standard test. The results are below:*

	Pass	*Fail*	*total*
Supplement	*284*	*216*	*500*
No supplement	*233*	*267*	*500*
total	*517*	*483*	

   (a) Test the hypothesis that results were independent of whether the subjects had taken the vitamin supplement at the 5% significance level.

   Under the independence assumption, the expected values for each class are 258.5, 241.5, 258.5 and 241.5, so the chi-squared statistic is $25.5^2\left(\frac{2}{258.5} + \frac{2}{241.5}\right) = 10.416$, which should be a chi-squared statistic with one degree of freedom. This has a p-value of 0.00125, so the hypothesis is rejected.

   (b) Does this show that the vitamin supplement causes individuals to get better scores in the tests. If not, give a possible alternative explanation.

Assuming that there was no problem with the choice of which subjects received the vitamin supplement (either if the subjects were chosen at random, or if some form of stratified sampling was performed) the result does show that the vitamin supplement causes the improved scores, since we know that the taking of the vitamin supplement is determined at random, and not by any other factor which could cause the better test scores.

9. *A scientist is investigating whether global warming is causing hurricanes in more regions. He looks the history of 200 cities, and whether they experienced hurricanes in the period 1990–2000 or in the period 2000–2010. The results are below:*

	Hurricanes 2000–2010	*No Hurricanes 2000–2010*	*total*
Hurricanes 1990–2000	*67*	*14*	*81*
No Hurricanes 1990–2000	*19*	*100*	*119*
total	*86*	*114*	

Test the hypothesis that the probability of a city experiencing a hurricane was the same for these two periods.

The probabilities are the same if and only if the probability of a city experiencing a hurricane in 1990–2000 but not 2000–2010 is the same as the probability of experiencing a hurricane in 2000–2010, but not 1990–2000. There are 14 cities in the first category, and 19 in the second. The relative log-likelihood of these having the same probability is therefore $33\log(16.5) - 14\log(14) - 19\log(19) = -0.380$. The chi-square statistic is twice this, or 0.761. This should be a chi-squared statistic with one degree of freedom, so it has a p-value of 0.383. Therefore, the hypothesis is not rejected.

10. *A coin is tossed 100 times. It comes up heads 39 times. Find a 10% significance region for the probability that the coin comes up heads, based on a chi-squared statistic* $\frac{(N-\mathbb{E}(N))^2}{\mathbb{E}(N)} + \frac{(N-\mathbb{E}(N))^2}{100-\mathbb{E}(N)}$.

For a given probability p, the expected number of heads is $100p$, so the chi-squared statistic is $(39-100p)^2\left(\frac{1}{100p} + \frac{1}{100(1-p)}\right)$. The 10% p-value for a chi-squared distribution with one degree of freedom is 2.70554345, so the 10% significance region is found by solving

$(39-100p)^2\left(\frac{1}{100p} + \frac{1}{100(1-p)}\right) = 2.70554345$

$(39-100p)^2\left((1-p)+p\right) = 270.554345p(1-p)$

$1521 - 7800p + 10000p^2 = 270.554345p - 270.554345p^2$

$1521 - 8070.554345p + 10270.554345p^2 = 0$

$p = \frac{8070.554345 \pm \sqrt{8070.554345^2 - 4\times 1521\times 10270.554345}}{2\times 10270.554345} = .4026 \pm .0792,$

so the 10% significance region is $[0.323, 0.482]$.

11. *(a) A coin is tossed 10 times. The probability of coming up heads is p. Show that the number of times heads comes up is a sufficient statistic for p.*

The likelihood is $L(p) = p^N(1-p)^{100-N}$, where N is the number of heads. This is clearly a function of the number of heads, so N is a sufficient statistic for p.

(b) What is the probability of the sequence HHTTHHHTHTT given that the number of heads is 5.

Given that the number of heads is 5, all $\binom{10}{5}$ sequences with exactly 5 heads are equally likely (regardless of p). Therefore the conditional probability of this sequence is $\frac{1}{\binom{10}{5}}$.

12. *The number of individuals taking an online survey follows a Poisson distribution with parameter 200. A particular question has two answers. Let the number of people giving each answer be A, and B. We want to estimate the probabilities of each answer.*

(a) Show that $A + B$ is an ancilliary statistic.

The likelihood is $e^{-200}200^{A+B}p^A(1-p)^B$. The maximum likelihood estimate of p is $\frac{A}{A+B}$. It is therefore clear that $(\hat{p}, A+B)$ is a minimal set of sufficient statistics. (Doubling both A and B does not affect $\hat{p}$ but does affect the likelihood function, so $\hat{p}$ is not sufficient.) So to show that $A+B$ is ancilliary, we just need to show that the distribution of $A+B$ does not depend on p. We know that $A+B$ is the number of people who take the survey, and has a Poisson distribution with parameter 200, whatever the value of p.

(b) In the case, $A = 89$, $B = 98$, use a conditional test, to calculate the significance of this result for the hypothesis that the probability of A is 0.6. [You may use a normal approximation to the Binomial distribution.]

Conditional on $A + B = 187$, under the hypothesis, the distribution of A, is binomial with $n = 187$ and $p = 0.6$. This can be approximated as a normal distribution with mean 112.2 and variance 44.88. Using a likelihood ratio test, the log-likelihood ratio is $89\log(112.2) - 89\log(89) + 98\log(74.8) - 98\log(98) = -5.858110703$. Trying a few values of A gives:

A	log-likelihood ratio
135	-6.068034369
132	-4.540840208
134	-5.532735634

This means that the significance of this result is the probability that $A \leqslant 89$ or $A \geqslant 135$. The probability of this is $1 - P(89.5 \leqslant A \leqslant 134.5)$. We use a normal approximation to get $1 - \left(\Phi\left(\frac{134.5-112.2}{\sqrt{44.88}}\right) - \Phi\left(\frac{89.5-112.2}{\sqrt{44.88}}\right)\right) = 2 - \Phi(3.34) - \Phi(3.39) = 0.0007$.

13. In two years, the results in a certain course at a university are:

Grade	*Year 1*	*Year 2*
Pass	*13*	*17*
Fail	*8*	*4*

Calculate the exact significance level for testing whether the probabilities of passing were the same in the two years. [Hint, the total number of results where the total number of students who pass during the two years is 11058116888.]

Under the hypothesis, the log-likelihood is $30\log(15)+12\log(6)-42\log(21) = -25.127322721$, while without the hypothesis, the log-likelihood is $13\log(13)+8\log(8)+17\log(17)+4\log(4)-42\log(21) = -24.180264111$. The log-likelihood ratio is therefore -0.94705861.

We condition on the total number of passes being 30, to get the results are less significant than the observed results only if between 14 and 16 students pass in the first year.

The probability of this is proportional to the number of outcomes with this condition. That is $\binom{21}{14}\binom{21}{16} + \binom{21}{15}^2 + \binom{21}{16}\binom{21}{14} = 7676945136$. The total number of outcomes with 30 passes in total across the two years is $\binom{42}{30} = 11058116888$.

The probability of getting a less significant outcome is therefore $\frac{7676945136}{11058116888} = 0.6942$, so the significance of the observed outcome is 0.3058.

14. The number of insurance claims paid on a certain policy is assumed to follow a binomial distribution with $n = 2$ and unknown p. The observed frequencies are :

Number	*Frequency*
0	*87*
1	*151*
2	*102*

Calculate the exact significance level, using a conditional test of the observed frequencies under the Binomial hypothesis.

Under the binomial hypothesis, the likelihood is $p^{\sum X_i}(1-p)^{2n-\sum X_i}$, so the sample mean $\hat{p} = \frac{\sum X_i}{n}$ is a sufficient statistic for p. Under the hypothesis, the log-likelihood is $325\log(1-p)+355\log(p)+151\log(2) = -366.013$, where $p = \frac{355}{680}$. Without the hypothesis, the log-likelihood is $87\log(87)+151\log(151)+102\log(102)-340\log(340) = -363.951$. The log-likelihood ratio is therefore -2.062.

To find the significance level, we condition on $\sum iX_i = 355$. We calculate the probability, conditional on this sum, that the log-likelihood ratio is at most -2.015.

Conditional on the sum, all the values are determined by the number of zeros, since the number of twos must be 15 more than the number of zeros.

Clearly, if the number of zeros is more than 87, then the result is more significant. Under the hypothesised distribution, the expected number of ones is slightly under 170, so similar levels of significance should occur when the number of ones is about 189, which makes the number of zeros 68. We compute the following:

no. of zeros	no. of ones	no. of twos	log-likelihood under binomial	log-likelihood	log-likelihood ratio
68	189	83	-339.67	337.46	2.21
69	187	84	-341.060	339.282	1.78

So we see that the values 69–86 are not as significant. To find the exact significance level, we need to find the probability that the number of such values is in this range, conditional on the value of $\hat{p}$.

Conditional on $\hat{p} = \frac{355}{680}$, the probability that there are n zeros is $\frac{P(Z=n,T=n+15)}{P\left(\hat{p}=\frac{355}{680}\right)}$

The numerator is $P(Z = n, T = n + 15) = \binom{340}{n,n+15,340-15-2n} p^{2n}(1 - p)^{2(n+15)} 2^{340-2n-15}(p(1-p))^{340-2n-15} = \binom{340}{n,n+15,325-2n} p^{325}(1-p)^{355} 2^{325-2n}$.

The denominator is the sum of these values over all n. This gives us a probability of 0.9499, so the significance level is 0.0500.